Faculty Working Papers

QUEUEING NETWORKS UNDER THE CLASS OF STATIONARY SERVICE POLICIES, II

Zvi Rosberg, Visiting Assistant Professor,
Department of Business Administration

#647

College of Commerce and Business Administration
University of Illinois at Urbana-Champaign

9. Schrage, L. E., "A Proof of the Optimality of the Shor
ing Processing Time Discipline," *Opns. Res.* 16, 687-69

FACULTY WORKING PAPERS

College of Commerce and Business Administration

University of Illinois at Urbana-Champaign

February 1980

QUEUEING NETWORKS UNDER THE CLASS OF STATIONARY SERVICE POLICIES, II

Zvi Rosberg, Visiting Assistant Professor,
Department of Business Administration

#647

Summary

In this study we apply the results obtained in Rosberg [2] for an exponential queueing network, with different classes of customers under the class of stationary service policies and with the mean cost per unit time as the loss function. A subclass of service policies which is used to reduce the loss function, as well as a heuristic service policy which is found to work well in the examples presented in the paper, are derived. An interactive computer program which uses the subclass of service policies and simulation for finding a service policy which yields a low value for the loss function is also presented. This program is applied to several examples of computer models.

In this paper we consider an exponential queueing network, with different classes of customers under the class of stationary service policies and the mean loss per unit time as the loss function. Here, we apply the results obtained in [2]. The model and the main results from [2] are given in section 1. In section 2, we derive a subclass of service policies (ψ-priority service policies) which is used to reduce the loss function. A heuristic service policy which is a ψ-priority policy and found to work well in the examples presented in section 3, is also given. Furthermore, we present an interactive computer program which uses the ψ-priority policies and simulation for finding a service policy which yields a low value of the loss function. The program is applied in section 3 to several examples of computer models.

1. THE MODEL, NOTATIONS AND EARLIER RESULTS

As in Rosberg [2] we consider an exponential queueing network with different classes of customers and cost for staying in the system, which is defined by the set of parameters

$$\Gamma = (A,B,\lambda,q_\alpha(\beta),\mu_\alpha(\beta),R(\beta),c_\alpha(\beta)\,|\,\alpha\varepsilon A,\beta\varepsilon B) \quad , \text{ where}$$

A denotes a finite set $\{1,2,\ldots,a\}$ of service stations serving customers independently and simultaneously. Each service station allows an unbounded queue.

B is a finite set $\{1,2,\ldots,b\}$ of classes of customers.

λ is the total arrival rate of customers from outside the system to all the stations. The arrival process is assumed to be Poisson.

$q_\alpha(\beta)$ is the probability that an arriving customer belongs to class β and joins to station α, $\alpha\varepsilon A$, $\beta\varepsilon B$ and $\sum_{\alpha,\beta} q_\alpha(\beta) = 1$.

$\mu_\alpha(\beta)$ is the service rate of customers of class β, $\beta\varepsilon B$, when they are provided service at the service station α, $\alpha\varepsilon A$. The service requirements are exponential r.v.'s, mutually independent and independent of the arrival process. (From [2], we may assume, without loss of generality, that $\mu_\alpha(\beta) = \mu$ for any $\alpha\varepsilon A$, $\beta\varepsilon B$.)

$R(\beta)$ is a sub-stochastic matrix, which describes the transition probabilities among the service stations of a customer of class β, $\beta\varepsilon B$. The (α,s) element of $R(\beta)$, denoted by $r_{\alpha s}(\beta)$, $\alpha,s\varepsilon A$, is the probability that a customer of class β, $\beta\varepsilon B$, who has been provided service at station α, will move next to station s. The probability that a customer will leave the system is $1 - \sum_s r_{\alpha s}(\beta)$.

$c_\alpha(\beta)$ is the cost of staying a unit of time at the service station α, $\alpha\varepsilon A$, for customers of class β, $\beta\varepsilon B$.

Let $B^* = B\cup\{0\}$, where 0 stands for one dummy customer present in each service station, whose parameters are $c_\alpha(0) = q_\alpha(0) = 0$, $\mu_\alpha(0) = \mu$ and $r_{\alpha\alpha}(0) = 1$, for any $\alpha\varepsilon A$.

To complete the definition of the queueing system, we must still define the service policy, i.e., a decision rule indicating which customer is served at each of the service stations at any

moment of time. Let $n = \{n_\alpha(\beta) | \alpha\varepsilon A, \beta\varepsilon B\}$ be the characterization, at any moment of time, of the queues in system Γ, where $n_\alpha(\beta)$ is the number of customers of class β at service station α at that moment. Furthermore, let any instant of time be a potential decision epoch. We consider any service policy which statisfies the following properties:

(i) At any moment of time the decision rule is a function of the state n only.

(ii) The servers are not allowed to be idel when there are customers at their stations. (I.e., the dummy customer, 0, is being served only when the queue is empty.)

(iii) The service of customers at each service station may be interrupted without losing any service duration which has already been provided.

The mean loss per unit time is taken as the loss function, L. From [2], section 2 we have that under stationary condition

$$L = (c,\overline{n}) \ ,$$

where $\overline{n} = \{\overline{n}_\alpha(\beta) | \alpha\varepsilon A, \beta\varepsilon B\}$ and $\overline{n}_\alpha(\beta)$ is the expected number of customers of class β at service station α under stationary conditions for a given service policy.

To ensure stability of the system, under any given service policy we make the following two assumptions about the relative traffic intensity at each service station.

ASSUMPTION 1. *For any* $\alpha \varepsilon A$, $\beta \varepsilon B$, *there exists an integer* $k \geq 1$, *such that*, $1 - \sum_{s \varepsilon A} r^k_{\alpha s}(\beta) > 0$, *where* $r^k_{\alpha s}(\beta)$ *are the elements of* $R^k(\beta)$, *the* $k^{\underline{th}}$ *power of* $R(\beta)$.

From assumption 1, it follows that, for any $\beta \varepsilon B$, there is a unique nonnegative solution $\lambda(\beta) = (\lambda_1(\beta), \lambda_2(\beta), \ldots, \lambda_a(\beta))$ to the equation

$$\lambda(\beta)(I - R(\beta)) = \lambda\, q(\beta) ,$$

where $q(\beta) = (q_1(\beta), q_2(\beta), \ldots, q_a(\beta))$ and I is the identity matrix of order a. (See [2], section 2.)

Let $\rho_\alpha(\beta) = \lambda_\alpha(\beta)/\mu$.

ASSUMPTION 2. *For any* $\alpha \varepsilon A$, $\sum_{\beta \varepsilon B} \rho_\alpha(\beta) < 1$.

For further results, we shall need the following notations and results appearing in Rosberg [2]. For any $i,k \varepsilon A$ and $j,m \varepsilon B^*$ let $y_{i,j}(k,m)$ be the expected number of customers of class j at station i, under stationary conditions, given that a customer of class m is provided service at station k, times the probability that a customer of class m is provided service at station k, under any given service policy. Define the matrices $Y = (y(i,j))$ and $T = (t(i,j))$, the column vectors $q = (q(i))$ and $c = (c(i))$ as follows:

$$y((m-1)a+k,(j-1)a+i)) = \begin{cases} y_{i,j}(k,m) & \text{if} \quad 1 \leq i,k \leq a \quad \text{and} \quad 1 \leq j,m \leq b , \\ y_{i,j}(k,0) & \text{if} \quad 1 \leq i,k \leq a ; 1 \leq j \leq b \quad \text{and} \quad m=b+1 , \\ 0 & \text{if} \quad 1 \leq i,k \leq a ; 1 \leq m \leq b+1 \quad \text{and} \quad j=b+1 , \end{cases}$$

$$t((j-1)a+i,(m-1)a+k) = \begin{cases} t(i,j,k,m) & \text{if } 1\le i,k\le a \text{ and } 1\le j,m\le b\,, \\ 0 & \text{if } 1\le i,k\le a; 1\le j\le b \text{ and } m=b+1\,, \\ 0 & \text{if } 1\le i,k\le a; 1\le m\le b+1 \text{ and } j=b+1\,, \end{cases}$$

where

$$t(i,j,k,m) = \begin{cases} 2\rho_i(j)(1-r_{ii}(j)) & \text{if } i=k \text{ and } j=m\le b\,, \\ -(\rho_i(m)r_{i,k}(m) + \rho_k(m)r_{k,i}(m)) & \text{if } i\ne k \text{ and } j=m\le b\,, \\ 0 & \text{otherwise}\,. \end{cases}$$

$$c((m-1)a+k) = \begin{cases} c_k(m) & \text{if } 1\le k\le a \text{ and } 1\le m\le b\,, \\ 0 & \text{if } 1\le k\le a \text{ and } m=b+1\,, \end{cases}$$

$$q((m-1)a+k) = \begin{cases} q_k(m) & \text{if } 1\le k\le a \text{ and } 1\le m\le b\,, \\ 0 & \text{if } 1\le k\le a \text{ and } m=b+1\,. \end{cases}$$

Let $\Omega = \{(\alpha,\beta)\,|\,\alpha\varepsilon A,\ \beta\varepsilon B^*\}$. Any $(\alpha,\beta)\varepsilon\Omega$ represents a phase of service in the system Γ. For any phase of service $(\alpha,\beta)\varepsilon\Omega$, define an order index

$$I_{(\alpha,\beta)} = \max_{\substack{M\subseteq A:\\ \alpha\varepsilon M}} \left\{ \frac{c_\alpha(\beta) - \sum_{s\notin M} r^{\beta}_{\alpha,s}(M)c_s(\beta)}{\gamma_{(\alpha,\beta)}(M)} \right\}, \tag{1}$$

(the maximum is taken over all subsets $M \subseteq A$ containing α), where $r^{\beta}_{\alpha,s}(M)$ is the probability that a customer of class β, starting service at station $\alpha\varepsilon M$, will enter first to station $s\notin M$ at his first exist from M and $\gamma_{(\alpha,\beta)}(M)$ is the mean total number of visits to stations in M of a customer of class β, starting service at station α, up to his first

exit from the set M. Note that the numerator in (1) is the expectation of the difference between the staying cost of a customer of class β who is present in station α and the staying cost of this customer immediately after leaving first the set M. It is clear that different phases (α,β), may attain the same value $I_{(\alpha,\beta)}$ and that $I_{(\alpha,0)} = 0$ for any $\alpha \varepsilon A$, which is the lowest value attained.

Define a complete order, $\prec$, henceforth "Klimov order" among the phases of service in Ω.

DEFINITION 1.

(i) For any (α,β), $(\alpha',\beta') \varepsilon \Omega$, $(\alpha,\beta) \prec (\alpha',\beta')$ iff $I_{(\alpha,\beta)} > I_{(\alpha',\beta')}$.

(ii) For any set of phases in $\{(\alpha,\beta) | \alpha\varepsilon A, \beta\varepsilon B\}$ *which obtain the same value in (1), the order* $\prec$*, is chosen arbitrarily.*

(iii) For the set of phases $\{(\alpha,0) | \alpha\varepsilon A\}$, $(\alpha,0) \prec (\alpha',0)$ iff $\alpha < \alpha'$.

From [1], it follows that the order on the set Ω, which is defined in [2], section 3, is the same as the order defined in definition 1 here.

Let $(\alpha,\beta)_1 \prec (\alpha,\beta)_2 \prec \ldots \prec (\alpha,\beta)_{(b+1)a}$, be the ordered phases in Ω. Rename the phases in Ω such that $(\alpha,\beta)_i \ \varepsilon \ \Omega$ will be denoted by i. With this notation the rows and the columns of Y, T, q and c will be permutated accordingly. For any $\alpha\varepsilon A$, $\beta\varepsilon B$, the row $(\beta-1)a+\alpha$ in Y, T, q and c will replace the row $(\alpha,\beta)_i$ and similarly for the columns of Y and T. Henceforth the notations $\Omega = \{1,2,\ldots,(b+1)a\}$, Y, T, q and c will be used to denote the appropriate reordered phases, matrices

For any i, $1 \leq i \leq (b+1)a$ let $\Omega_i = \{1,2,\ldots,i\}$. Further, for any $p \in \Omega_i$ let $\gamma_p(\Omega_i)$ be the mean total number of visits to phases of service in Ω_i, of a customer starting with phase p, up to his first exit from the set Ω_i. Furthermore, for any $p \in \Omega_i$ define recursively the elements $c_p(\Omega_i)$ in the following way:

$$c_p(\Omega_{(b+1)a}) = c(p) \ , \quad \text{for any} \quad p \in \Omega_{(b+1)a} = \Omega \ ;$$

$$c_p(\Omega_{i-1}) = c_p(\Omega_i) - \gamma_p(\Omega_i)\,\frac{c_i(\Omega_i)}{\gamma_i(\Omega_i)} \quad \text{for any} \quad p \in \Omega_{i-1} \text{ and } i,\ 1 < i \leq (b+1)a.$$

Let,

$$u_p = (\gamma_1(M_p), \gamma_2(M_p), \ldots, \gamma_p(M_p), 0, \ldots, 0)'$$

and

$$v_p = (0,0,\ldots,0,1,1,\ldots,1)' \ ,$$

where u_p, v_p have $(b+1)a$ elements, v_p has p zeros preceding the one's and v' is the transpose of the vector v.

Further, let

$$z_p = c_p(M_p)/\gamma_p(M_p) \ ,$$

$$h_p = (v_p, Yu_p)/(1-\lambda(q,u_p)/a\mu) \ ,$$

$$g_p = (Tu_p, u_p)/2(1-\lambda(q,u_p)/a\mu) \ , \qquad (2)$$

$$z = (z_1, z_2, \ldots, z_{ab})' \ , \ h = (h_1, h_2, \ldots, h_{ab})'$$

and

$$g = (g_1, g_2, \ldots, g_{ab})' \ ,$$

where $(x,y) = \sum_{i=1}^{n} x_i y_i$, for any two vectors $x = (x_1, x_2, \ldots, x_n)$ and $y = (y_1, y_2, \ldots, y_n)$.

Still further, let

$$LB1 = (z,g)/a \ ,$$

LB2 be the value obtained in [2], section 3, by solving the appropriate linear programming problem defined there,

$$LB3 = \sum_{\beta\varepsilon B} (\min_{i\varepsilon A} c_i(\beta)) \sum_{\alpha\varepsilon A} \rho_\alpha(\beta)/(1 - \rho_\alpha(\beta)) \ ,$$

$$UB = \sum_{\substack{\alpha\varepsilon A \\ \beta\varepsilon B}} c_\alpha(\beta)\rho_\alpha(\beta)/(1 - \sum_{j\varepsilon B} \rho_\alpha(j))$$

and L_0 be the value of the loss function L, under the optimal service policy. (Note that UB is the value of L under the Service-Sharing service policy. See [2], section 3.) The main results obtained in [2] are given in the following two theorems.

THEOREM 1. *Under any service policy*

$$L = (z,g)/a + (z,h)/a \ .$$

In addition, $z \geq 0$ *and* $h \geq 0$.

THEOREM 2.

$$UB \geq L_0 \geq \max\{LB1, LB2, LB3\} \ .$$

2. REDUCING THE LOSS FUNCTION

From (2) and theorem 1 it follows that the loss function L will be reduced if and only if (z,h) will be reduced.

From (2) we have

$$(z,h) = \sum_{\substack{p,p'\in\Omega \\ p'\prec p}} w_{p'}(p)y(p,p') \, , \tag{3}$$

where $y(p,p')$ are the elements of the matrix Y and

$$w_{p'}(p) = \sum_{k=p'}^{ab} \left(c_k(M_k)\gamma_{p'}(M_k)/\gamma_k(M_k)(1-\lambda(q,u_k)/a\mu)\right) \times I(k<p) \times I(p'\le ab) \, . \tag{4}$$

Here, $I(x < y) = \begin{cases} 1 & x < y \, , \\ 0 & x \geq y \, . \end{cases}$

Note that from theorem 1 we have $w_{p'}(p) \geq 0$.

Using the original notation for the phases of service p, p' in (3) we have

$$(z,h) = \sum_{(\alpha',\beta')\prec(\alpha,\beta)} w_{(\alpha',\beta')}(\alpha,\beta)y_{\alpha',\beta'}(\alpha,\beta) \, .$$

Furthermore, from (4) it follows that for any phases of service (α,β), (α',β'), $\alpha,\alpha'\varepsilon A$, $\beta,\beta'\varepsilon B^*$

$$w_{(\alpha',\beta')}(i,j) \geq w_{(\alpha',\beta')}(\alpha,\beta) \quad \text{for any} \quad (\alpha,\beta) \prec (i,j) \, .$$

The expressions $w_{(\alpha',\beta')}(\alpha,\beta)$ are indpendent of the service policy chosen in system Γ. Thus, they can be considered as weights of the expectations $y_{\alpha',\beta'}(\alpha,\beta)$ which do depend on the service policy.

It is easy to check that for systems with more than one service station, none of the service policies can reduce (z,h) to zero.

To improve a service policy we need to reduce some of the expectations $y_{\alpha',\beta'}(\alpha,\beta)$, but in doing so some others are increased. Thus, the goal will be to find service policies which reduce the expectations $y_{\alpha',\beta'}(\alpha,\beta)$ with higher weights $w_{(\alpha',\beta')}(\alpha,\beta)$ at the expense of those with lower weights.

Let $n = \{n_\alpha(\beta) | \alpha\varepsilon A, \beta\varepsilon B\}$ be any state of the system at any given, but arbitrary, moment of time t. In order to define a service policy we need to indicate the class of the customer which is served at station α, for any $\alpha\varepsilon A$ at this state n.

By providing service at station α to a customer from a specific class β, whenever the system is in state $n = \{n_{\alpha'}(\beta') | \alpha'\varepsilon A, \beta'\varepsilon B\}$ (at any moment of time t, independent of the specific moment t), the values $n_{\alpha'}(\beta')$, $\alpha'\varepsilon A$, $\beta'\varepsilon B$ determine in the long run the expectations $y_{\alpha',\beta'}(\alpha,\beta)$. Obviously, by choosing the class to be served at any station $\alpha\varepsilon A$, we specify which of the expectations $y_{\alpha',\beta'}(\alpha,\beta)$ will be affected.

In order to attain the goal mentioned above, it follows from (5) and (6), that we must choose the customers for service in the following way:

At any state of the system $n = \{n_{\alpha'}(\beta') | \alpha'\varepsilon A, \beta'\varepsilon B\}$, for any station $\alpha\varepsilon A$ and for any two classes of customers $j,m\varepsilon B$ such that $(\alpha,m) \prec (\alpha,j)$,

a customer of class j will be preferred for service to a customer of class m, whenever the values $n_{\alpha'}(\beta')$ for $(\alpha',\beta') \prec (\alpha,j)$, are relatively small. If they are not, a customer of class m will be preferred to a customer of class j.

This preference of customers can be obtained by a parameterized class of service policies defined as follows.

For any $\alpha\varepsilon A$, $\beta\varepsilon B$ let $\psi(\alpha,\beta) = \{\psi_{(\alpha',\beta')}(\alpha,\beta) | \alpha'\varepsilon A, \beta'\varepsilon B$ and $(\alpha',\beta') \prec (\alpha,\beta)\}$, where $\psi_{(\alpha',\beta')}(\alpha,\beta)$ are nonnegative integers and for any $\alpha\varepsilon A$, $j,m\varepsilon B$ such that $(\alpha,m) \prec (\alpha,j)$,

$$\psi_{(\alpha',\beta')}(\alpha,j) \leq \psi_{(\alpha',\beta')}(\alpha,m) \quad \text{for any} \quad \alpha'\varepsilon A, \beta'\varepsilon B \; . \qquad (7)$$

The set $\psi(\alpha,\beta)$ will define a service domain for customers of class β at station α.

DEFINITION 2.

For any $\alpha\varepsilon A$, $\beta\varepsilon B$ *and any set* $\psi(\alpha,\beta)$, *a customer of class* β *present at station* α *is in the* $\psi(\alpha,\beta)$*-service domain when the system is in state* $n = \{n_{\alpha'}(\beta') | \alpha'\varepsilon A, \beta'\varepsilon B\}$, *iff*

$$n_{\alpha'}(\beta') \leq \psi_{(\alpha',\beta')}(\alpha,\beta) \quad \textit{for any} \quad (\alpha',\beta') \prec (\alpha,\beta) \; .$$

Let $\psi = \{\psi(\alpha,\beta) | \alpha\varepsilon A, \beta\varepsilon B\}$. Any set ψ, satisfying (7) is a set of parameters which will determine a service policy in the class of service policies defined below.

For any set of parameters ψ, state of the system n and service station $\alpha\varepsilon A$, let

$B^n_\alpha(\psi) = \{ \beta \mid \beta\varepsilon B, n_\alpha(\beta) > 0$ and customer of class β is in the $\psi(\alpha,\beta)$-service domain.$\}$,

$$B^n_\alpha = \{ \beta \mid \beta\varepsilon B^* \text{ and } n_\alpha(\beta) > 0\} \supseteq \{(\alpha,0)\} \neq \phi . \qquad (8)$$

For any set ψ, we shall define a ψ-priority service policy.

DEFINITION 3.

At any state n *and at each of the service stations* $\alpha\varepsilon A$, *the customer which is being served at station* α *according to the* ψ*-priority service policy is any customer of class* β_α, *where* β_α *is determined as follows:*

(i) If $B^n_\alpha(\psi) \neq \phi$, *then* β_α *is the class which satisfies*

$$\beta_\alpha \varepsilon B^n_\alpha(\psi)$$

and

$$(\alpha,\beta_\alpha) > (\alpha,\beta) \quad \textit{for any} \quad \beta\varepsilon B^n_\alpha(\psi) .$$

(ii) If $B^n_\alpha(\psi) = \phi$, *then* β_α *is the class which satisfies*

$$\beta_\alpha \varepsilon B^n_\alpha$$

and

$$(\alpha,\beta_\alpha) \prec (\alpha,\beta) \quad \textit{for any} \quad \beta \varepsilon B^n_\alpha .$$

For any station $\alpha\varepsilon A$, if there are customers in the phases of service (α,β), $\beta\varepsilon B$, which are in the $\psi(\alpha,\beta)$-service domain, then the ψ-priority policy preferred for service the highest "Klimov

ordered" phase among them. If there are no customers in their service domain, then the lowest "Klimov ordered" nonempty phase is preferred.

The problem in hand now, is how to select a set of parameters ψ, which will yield a low value for the loss function L. Analytically, this is extremely difficult because of the complicated form of the service policies involved. However, this can be done by using a simulation of the model. For that purpose we wrote an interactive computer program, programmed in the APL programming language and running in an IBM/370 computer machine.

The program has two major parts. In the first part the program performs:

(i) A data entry of the system parameters Γ by the user.

(ii) A computation of the intensities $\rho_\alpha(\beta)$, $\alpha\varepsilon A$, $\beta\varepsilon B$, which are defined in section 1.

(iii) A computation of the bounds LB1, LB2, LB3 and UB which are defined in section 1.

(iv) A computation of the "Klimov order" among the phases of service (α,β), $\alpha\varepsilon A$, $\beta\varepsilon B$.

(v) A computation of the weights $w_{(\alpha',\beta')}(\alpha,\beta)$, $\alpha,\alpha'\varepsilon A$, $\beta,\beta'\varepsilon B$ which are defined in (4).

In the second part the program performs:

(i) A simulation of the system Γ under any given stationary service policy, which may be entered or altered by the user at any moment during the program life.

(ii) A computation of the values $\rho_\alpha^T(\beta)$, L^T and $y^T_{\alpha',\beta'}(\alpha,\beta)$ at any moment T of the simulation, where

$$\chi_{\alpha,\beta}(t) = \begin{cases} 1 & \text{if a customer of class } \beta \text{ is being served at station } \alpha \text{ at time } t \text{ of the simulation ,} \\ 0 & \text{otherwise ,} \end{cases}$$

$$\rho_\alpha^T(\beta) = \frac{1}{T}\int_0^T \chi_{\alpha,\beta}(t)dt \ ,$$

$$L^T = \frac{1}{T}\int_0^T \sum_{\substack{\alpha\varepsilon A \\ \beta\varepsilon B}} c_\alpha(\beta) n_\alpha^t(\beta)dt \ .$$

Here, $n_\alpha^t(\beta)$ is the number of customers of class β at station α, at time t of the simulation.

$$y^t_{\alpha',\beta'}(\alpha,\beta) = \begin{cases} n_{\alpha'}(\beta') & \text{if at time } t \text{ of the simulation, a customer of class } \beta \text{ is being served at station } \alpha \text{ and } n^t_{\alpha'}(\beta') = n_{\alpha'}(\beta'), \\ 0 & \text{otherwise,} \end{cases}$$

$$y^T_{\alpha',\beta'}(\alpha,\beta) = \frac{1}{T}\int_0^T y^t_{\alpha',\beta'}(\alpha,\beta)dt \ .$$

Under stationary conditions we have

$$L^T \xrightarrow[T\to\infty]{} L \ ; \ y^T_{\alpha',\beta'}(\alpha,\beta) \xrightarrow[T\to\infty]{} y_{\alpha',\beta'}(\alpha,\beta)$$

and in addition, from theorem 3.1.2 in [2] we have

$$\rho_\alpha^T(\beta) \xrightarrow[T\to\infty]{} \rho_\alpha(\beta) \ .$$

The stopping time of the simulation, under a given service policy, is determined from the convergence of the values $\rho_\alpha^T(\beta)$, $\alpha\epsilon A$, $\beta\epsilon B$, to the values $\rho_\alpha(\beta)$, $\alpha\epsilon A$, $\beta\epsilon B$, which are found analytically in the the first part of the program and from the convergence of the sequence L^T.

The parameters $\psi_{(\alpha',\beta')}(\alpha,\beta)$, for the ψ-priority service policy under which the simulation of the model will run, are always chosen with respect to the values $w_{(\alpha',\beta')}(\alpha,\beta)$ so that higher values of $w_{(\alpha',\beta')}(\alpha,\beta)$ impose lower values on $\psi_{(\alpha',\beta')}(\alpha,\beta)$. There are not as many practical possibilities of choosing the set ψ as it seems, since in ergodic systems (see [2], section 2), the value of the loss function L, is mainly affected by the rules of the service policy at the states n, with small numbers of customers. (How small depends on the intensities $\rho_\alpha(\beta)$.)

An improvement of a chosen ψ-priority service policy is made by trial and error after any simulation. This improvement is based on the values L^T, $y^T_{\alpha',\beta'}(\alpha,\beta)$ obtained at the stopping time of the simulations under all the previous service policies.

Define two measures for evaluating a service policy which provide a loss L^T at the stopping time of the simulation,

$$E_f = \max\{LB1, LB2, LB3\}/L^T ,$$

$$I_m = (UB-L^T)/(L^T-\max\{LB1, LB2, LB3\}) ,$$

where the bounds LB1, LB3 and UB are calculated from the values $\rho_\alpha^T(\beta)$ rather than from the theoretical values $\rho_\alpha(\beta)$. The measure $0 \leq E_f \leq 1$ is an efficiency measure. The measure $0 \leq I_m < \infty$ is an improvement measure in comparison to the service-sharing policy.

We terminate the improvement process of the service policies when E_f is satisfactory enough (e.g., above 0.95) or when the simulations results do not leave possibilities of further improvement using the class of ψ-priority service policies.

In the examples that we analyzed, some of which are presented in section 3 below, 2-5 simulations were enough to determine a satisfactory service policy.

REMARK. We don't have as yet an algorithm which determines the best policy in the class of ψ-priority policies. Individual consideration, using the guidelines above, should be given for each particular system Γ.

We shall employ the similarity between the system Γ and Klimov's model (see [1]), to obtain a heuristic service policy which provides the best results in all the examples that we have analyzed. Consider the original set of the system parameters Γ, in another form, i.e.,

$$(\Omega,\ \lambda,\ q,\ \mu,\ R,\ c)\ ,$$

where $\Omega = \{(\alpha,\beta) \mid \alpha\varepsilon A, \beta\varepsilon B\}$ is the set of phases of service, q, c are the vectors defined in section 1, λ, μ are the arrival and service

$$R = \begin{bmatrix} R(1) & & & 0 \\ & R(2) & & \\ & & \ddots & \\ 0 & & & R(b) \end{bmatrix}.$$

Now, instead of having b service stations each of which provide service simultaneously to another set of phases, suppose there is only one server providing service at any moment of time to one phase only. Denote this revised system by Γ'.

The differences between the system Γ' and Klimov's model are: (i) In Klimov's model, only non-preemptive service policies are allowed and in the system Γ' preemption resume policies are also allowed. (ii) The service requirements in Γ' are exponential rather than general as in Klimov's model.

However, the optimal service policy in Γ' is the same as the optimal service policy in Klimov's model. (This can be seen in the same way as Klimov's model was analyzed.)

Thus from [1], the optimal service policy in Γ' is: if $I_{(\alpha,\beta)} > I_{(\alpha',\beta')}$, then the phase (α,β) has preference to receive service over phase (α',β'); but if $I_{(\alpha,\beta)} = I_{(\alpha',\beta')}$, the service order among (α,β) and (α',β') may be chosen arbitrarily.

The similarity between the system Γ' and the original system Γ, described above and the optimal service policy in system Γ', suggests the following heuristic policy which will be referred as "Klimov policy."

DEFINITION 4

At any state n *and at each of the service stations* $\alpha\varepsilon A$, *the customer which is being served at station* α *according to "Klimov policy" is any customer of class* $\beta_\alpha \varepsilon B_\alpha^n$, *where* B_α^n *is defined in (8) and* β_α *is any class which satisfies* $I_{(\alpha,\beta_\alpha)} \geq I_{(\alpha,\beta)}$ *for any* $\beta \varepsilon B_\alpha^n$.

Note that "Klimov policy" is a special case of a ψ-priority policy, when $\psi_{(\alpha',\beta')}(\alpha,\beta) = 0$ for any $\alpha,\alpha'\varepsilon A$ and $\beta,\beta'\varepsilon B$.

3. EXAMPLES

We shall apply the method and the computer program presented in section 2, to several examples of computer models. In each example, any service station is one of two types, a Central Processing Unit (CPU) or a Data Transmission Unit (DTU). The customers are the programs running in the computer, which are classified according to their service requirements. In our examples, there are three possible classes of customers. Class 1 is a class of CPU bounded customers. That is, programs for which the service requirement of each program from the CPU, once the program is in the CPU, is relatively higher than the service requirement from the DTU, once the program is in the DTU. Class 2 is a class of DTU bounded customers, which is defined in the same manner as class 1, but the relative amount of service in the CPU and the DTU is reversed. Class 3 is a class of CPU and DTU bounded customers. That is, programs, which have high demand of service from the CPU and the DTU, once they are there.

EXAMPLE 1.

We consider two service stations, station number 1 is the CPU and station number 2 is the DTU. Two classes of programs are running through the servers, class 1 consists of CPU bounded programs and class 2 consists of DTU bounded programs. The staying cost of all the customers are equal. Such a system is modeled by the following set of parameters Γ.

$$A = \{1,2\} \ ; \ B = \{1,2\} \ ; \ \lambda = 0.1 \ ; \ \mu = 1 \ ;$$

$$q_1(1) = q_1(2) = 0.5 \ ; \ q_2(1) = q_2(2) = 0 \ ; \ c_\alpha(\beta) = 1 \text{ for any } \alpha\varepsilon A, \ \beta\varepsilon B.$$

$$R(1) = \begin{bmatrix} 0.6 & 0.1 \\ 1 & 0 \end{bmatrix} \ ; \ R(2) = \begin{bmatrix} 0 & 0.6 \\ 0.2 & 0.8 \end{bmatrix} .$$

The results are summarized in the following tables.

The intensities and Klimov order are given in Table 1.1

The phases (α,β), $\alpha\varepsilon A$, $\beta\varepsilon B$	(1,1)	(1,2)	(2,1)	(2,2)	(1,0)	(2,0)
$\rho_\alpha(\beta)$	0.16667	0.125	0.01667	0.375	0.70833	0.60833
Klimov order rank	2	1	3	4	5	6

Table 1.1

The row of Klimov order ranks means that $(1,2) \prec (1,1) \prec (2,1) \prec$ $\prec (2,2) \prec (1,0) \prec (2,0)$ and $(1,0)$, $(2,0)$ are phases where station 1 and station 2 respectively are empty.

The weights $w_{(\alpha',\beta')}(\alpha,\beta)$ are given by the appropriate elements in table 1.2.

(α,β) \ (α',β')	(1,2)	(1,1)	(2,1)	(2,2)
(1,2)	0	0	0	0
(1,1)	.1026	0	0	0
(2,1)	.1965	.2348	0	0
(2,2)	.3636	.8476	.7799	0
(1,0)	1.3763	1.2189	1.2524	1.5190
(2,0)	1.3763	1.2189	1.2524	1.5190

Table 1.2

The analytical bounds and the simulation results are given in table 1.3.

	UB	L	LB1	LB2	LB3	E_f	I_m	T
Analytical values	1.0556	--	0.4828	0.917	0.9598	--	--	--
Analytical values under the service-sharing policy	--	1.0556	0.4828	0.917	0.9598	0.9093	0	--
Values from simulation under policy 1	1.0488	0.9960	0.4819	0.917	0.9540	0.95	1.25	2867.5
Values from simulation under policy 2	1.1168	1.0243	0.4993	0.917	1.0141	0.99	9.06	20868.5

Table 1.3

T is the stopping time of the simulation and policies 1 and 2 are ψ-priority service policies, where the parameters $\psi_{(\alpha',\beta')}(\alpha,\beta)$ are given by the appropriate elements in tables 1.4 and 1.5

(α,β) \ (α',β')	(1,2)	(1,1)	(2,1)	(2,2)
(1,2)	∞	∞	∞	∞
(1,1)	0	∞	∞	∞
(2,1)	3	2	∞	∞
(2,2)	2	1	1	∞

Table 1.4 (Policy 1)

(α,β) \ (α',β')	(1,2)	(1,1)	(2,1)	(2,2)
(1,2)	∞	∞	∞	∞
(1,1)	0	∞	∞	∞
(2,1)	0	0	∞	∞
(2,2)	0	0	0	∞

Table 1.5 (Policy 2)

The best results were obtained with policy 2, which the reader can recognize as the heuristic service policy. The evaluation measures for it are $E_f = .99$ and $I_m = 9.06$. Policy 2 gives absolute priority in the DTU to the CPU bounded customers and an absolute priority in

the CPU to the DTU bounded customers. This policy seems very reasonable when the staying costs of the customers are the same.

EXAMPLE 2.

Let us change slightly the system given in example 1 to a system where the staying cost of customers of class 2 is five times more expensive than the staying cost of customers of class 1. That is, $c_\alpha(1) = 1$, $c_\alpha(2) = 5$ for any $\alpha\varepsilon A$. The heuristic service policy used in this case, gives absolute priority to customers of class 2 at both of the service stations. This heuristic policy, with $E_f = 0.97$, again gives the best results in this example. Moreover, it also provides the best results when we increase λ from 0.1 to 0.2.

EXAMPLE 3.

We consider two service stations; station number 1 is the CPU and station number 2 is the DTU. There are three classes of customers; class 1 is a class of CPU bounded customers, class 2 is a class of DTU bounded customers and class 3 is a class of CPU and DTU bounded customers. The staying cost of all the customers are the same. The parameters of the system are:

$$A = \{1,2\} \; ; \; B = \{1,2,3\} \; ; \; \lambda = 0.3 \; ; \; \mu = 1$$

$$q_1(\beta) = 1/3 \; , \; q_2(\beta) = 0 \; , \; c_\alpha(\beta) = 1 \quad \text{for any} \quad \alpha\varepsilon A \quad \text{and} \quad \beta\varepsilon B \; .$$

$$R(1) = \begin{bmatrix} 0.6 & 0.1 \\ 1 & 0 \end{bmatrix} \; ; \; R(2) = \begin{bmatrix} 0 & 0.6 \\ 0.2 & 0.8 \end{bmatrix} \; ; \; R(3) = \begin{bmatrix} 0.6 & 0.1 \\ 0.4 & 0.6 \end{bmatrix} \; .$$

The intensities and Klimov order are given in table 3.1.

The phases (α,β)	(1,1)	(1,2)	(1,3)	(2,1)	(2,2)	(2,3)	(1,0)	(2,0)
$\rho_\alpha(\beta)$	0.3333	0.25	0.3333	0.0333	0.75	0.0833	0.0834	0.1334
Klimov order rank	3	1	2	4	6	5	7	8

Table 3.1

The weights $w_{(\alpha',\beta')}(\alpha,\beta)$ are given in table 3.2.

(α,β) \ (α',β')	(1,2)	(1,3)	(1,1)	(2,1)	(2,3)	(2,2)
(1,2)	0	0	0	0	0	0
(1,3)	0.1053	0	0	0	0	0
(1,1)	0.1053	0	0	0	0	0
(2,1)	0.2277	0.3061	0.3061	0	0	0
(2,3)	0.3279	0.5565	0.6734	0.4675	0	0
(2,2)	0.4771	1.1784	1.2207	1.1640	0.9950	0
(1,0)	6.6310	3.7425	3.4771	4.0358	5.0975	9.2307
(2,0)	6.6310	3.7425	3.4771	4.0358	5.0975	9.2307

Table 3.2

The analytical bounds and the simulation results are given in table 3.3.

	UB	L	LB1	LB2	LB3	E_f	I_m	T
Analytical values	17.5	--	5.7734	11.594	4.4587	--	--	--
Analytical values under service-sharing policy	--	17.5	5.7734	11.594	4.4587	0.66	0	--
From simulation under policy 1	18.39	17.65	5.6830	11.594	4.2091	0.65	0.12	5077.6
From simulation under policy 2	16.52	11.97	5.6567	11.594	4.2530	0.97	11.97	9201.6

Table 3.3

Policy 2 is the heuristic service policy and policy 1 is a ψ-priority policy, where the parameters $\psi_{(\alpha',\beta')}(\alpha,\beta)$ are given in table 3.4.

(α,β) \ (α',β')	(1,2)	(1,3)	(1,1)	(2,1)	(2,3)	(2,2)
(1,2)	∞	∞	∞	∞	∞	∞
(1,3)	0	∞	∞	∞	∞	∞
(1,1)	0	0	∞	∞	∞	∞
(2,1)	3	2	2	∞	∞	∞
(2,3)	2	1	1	1	∞	∞
(2,2)	0	0	1	0	0	∞

Table 3.4 (Policy 1)

The heuristic service policy with $E_f = 0.97$ gives the best results. The most preferred customers in the CPU, by the heuristic

policy are the DTU bounded customers and the least preferred are the CPU bounded customers. In the DTU, the preference order by the heuristic policy is reversed.

EXAMPLE 4.

In this example there are four service stations; station number 1 is the CPU and stations number 2, 3, 4 are DTU's, where stations numbers 1 and 2 are Channel + Disk devices and station 3 is a Channel + Tape device. The classes of customers are as in example 3, where customers of classes 1 and 2 use Disk devices only and customers of class 3 use the Tape device only. The parameters of the system are:

$$A = \{1,2,3,4\} \ ; \ B = \{1,2,3\} ; \lambda = 0.2 \ ; \ \mu = 1$$

$$q_1(\beta) = 1/3 \ ; \ q_2(\beta) = 0 \quad \text{for any} \quad \beta \varepsilon B \quad \text{and} \quad \alpha, \ 2 \leq \alpha \leq 4 \ ;$$

$$c_\alpha(1) = 1 \ ; \ c_\alpha(2) = 4 \ ; \ c_\alpha(3) = 2 \quad \text{for any} \quad \alpha \varepsilon A \ ;$$

$$R(1) = \begin{bmatrix} 0.6 & 0.05 & 0.05 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{bmatrix} \ ; \ R(2) = \begin{bmatrix} 0 & 0.3 & 0.3 & 0 \\ 0.2 & 0.8 & 0 & 0 \\ 0.2 & 0 & 0.8 & 0 \\ 0.2 & 0 & 0 & 0.8 \end{bmatrix}$$

$$R(3) = \begin{bmatrix} 0.6 & 0 & 0 & 0 \\ 0.2 & 0.8 & 0 & 0 \\ 0.2 & 0 & 0.8 & 0 \\ 0.2 & 0 & 0 & 0.8 \end{bmatrix}$$

The intensities and Klimov order are given in table 4.1.

The phases (α,β)	(1,1)	(1,2)	(1,3)	(2,1)	(2,2)	(2,3)	(3,1)	(3,2)	(3,3)	(4,1)	(4,2)	(4,3)	(1,0)	(2,0)	(3,0)	(4,0)
$\rho_t(\beta)$	0.222	0.167	0.222	0.011	0.25	0	0.011	0.25	0	0	0	0.111	0.389	0.739	0.739	0.889
Klimov order rank	3	1	2	9	6	12	7	4	11	8	5	10	13	14	15	16

Table 4.1

The weights $w_{(\alpha',\beta')}(\alpha,\beta)$ are given in table 4.2.

(α,β) \ (α',β')	(1,2)	(1,3)	(1,1)	(3,2)	(4,2)	(2,2)	(3,1)	(4,1)	(2,1)	(4,3)	(3,3)	(2,3)
(1,2)	0	0	0	0	0	0	0	0	0	0	0	0
(1,3)	1.017	0	0	0	0	0	0	0	0	0	0	0
(1,1)	1.335	0.796	0	0	0	0	0	0	0	0	0	0
(3,2)	1.372	0.889	0.096	0	0	0	0	0	0	0	0	0
(4,2)	1.372	0.889	0.096	0	0	0	0	0	0	0	0	0
(2,2)	1.372	0.889	0.096	0	0	0	0	0	0	0	0	0
(3,1)	2.07	1.063	0.267	1.047	1.047	1.047	0	0	0	0	0	0
(4,1)	2.07	1.063	0.267	1.047	1.047	1.047	0	0	0	0	0	0
(2,1)	2.07	1.063	0.267	1.047	1.047	1.047	0	0	0	0	0	0
(4,3)	2.266	1.112	0.339	1.341	1.341	1.341	0.091	0.091	0.091	0	0	0
(3,3)	2.266	1.112	0.339	1.341	1.341	1.341	0.091	0.091	0.091	0	0	0
(2,3)	2.266	1.112	0.339	1.341	1.341	1.341	0.091	0.091	0.091	0	0	0
(1,0)	5.169	2.564	1.403	5.695	5.695	5.695	1.446	1.446	1.446	2.903	2.903	2.903
(2,0)	5.169	2.564	1.403	5.695	5.695	5.695	1.446	1.446	1.446	2.903	2.903	2.903
(3,0)	5.169	2.564	1.403	5.695	5.695	5.695	1.446	1.446	1.446	2.903	2.903	2.903
(4,0)	5.169	2.564	1.403	5.695	5.695	5.695	1.446	1.446	1.446	2.903	2.903	2.903

Table 4.2

The analytical bounds and the simulation results are given in table 4.3.

	UB	L	LB1	LB2	LB3	E_f	I_m	T
Analytical values	6.415	--	1.236	4.718	4.596	--	--	--
Analytical values under service-sharing policy	--	6.415	1.236	4.718	4.596	0.73	0	--
From simulation under policy 1	6.791	5.921	1.241	4.718	4.611	0.79	0.72	12583.1
From simulation under policy 2	6.344	5.439	1.232	4.718	4.599	0.85	1.2	4392

<u>Table 4.3</u>

Policy 2 is the heuristic service policy and policy 1 is a ψ-priority policy, where the parameters $\psi_{(\alpha',\beta')}(\alpha,\beta)$ are given in table 4.4.

(α,β) \ (α',β')	(1,2)	(1,3)	(1,1)	(3,2)	(4,2)	(2,2)	(3,1)	(4,1)	(2,1)	(4,3)	(3,3)	(2,3)
(1,2)	∞	∞	∞	∞	∞	∞	∞	∞	∞	∞	∞	∞
(1,3)	0	∞	∞	∞	∞	∞	∞	∞	∞	∞	∞	∞
(1,1)	2	1	∞	∞	∞	∞	∞	∞	∞	∞	∞	∞
(3,2)	2	3	5	∞	∞	∞	∞	∞	∞	∞	∞	∞
(4,2)	2	3	5	∞	∞	∞	∞	∞	∞	∞	∞	∞
(2,2)	1	3	5	∞	∞	∞	∞	∞	∞	∞	∞	∞
(3,1)	1	3	4	3	3	3	∞	∞	∞	∞	∞	∞
(4,1)	1	3	4	3	3	3	∞	∞	∞	∞	∞	∞
(2,1)	1	3	4	3	3	3	∞	∞	∞	∞	∞	∞
(4,3)	1	3	4	3	3	3	5	5	5	∞	∞	∞
(3,3)	1	3	4	3	3	3	5	5	5	∞	∞	∞
(2,3)	1	3	4	3	3	3	5	5	[illegible]	∞	∞	∞

<u>Table 4.4</u> (Policy 1)

The heuristic service policy with $E_f = 0.85$ gives the best results. The most preferred customers in the CPU, by this policy are customers of class 2 (with the highest staying cost) and the least preferred are the customers of class 1 (with the lowest staying cost).

In all the DTU stations, customers of class 2 are still the most preferred and customers of class 3 are the least preferred.

From the examples presented above and from other examples that we analyzed, we found

a) The heuristic service policy always gave the best results.

b) Neither of the bounds LB2, LB3 is better. (In [2] we show that LB2 $\geq$ LB1, but LB1 is easier to compute.)

c) In the simulation process, the quality of a service policy can be detected right at the beginning.

ACKNOWLEDGMENT

Most of the results appearing in this paper are part of the author's Ph.D. dissertation which was carried out at the Hebrew University of Jerusalem, under the supervision of Uri Yechiali and Shmuel Zamir to whom I am most grateful.

REFERENCES

1. Klimov, G. P., "Time-Sharing Service Systems, II," *Theory of Prob. and Its Appl.* 23, 2, 314-321 (1978) (English translation).

2. Rosberg, Z., "Queueing Networks Under the Class of Stationary Service Policies, I," Working Paper No. 646, College of Commerce and Business Administration, University of Illinois at Urbana-Champaign, February (1980).